

# Multilevel Sequential Monte Carlo Samplers for Normalizing Constants

Pierre Del Moral\*   Ajay Jasra†   Kody Law‡   Yan Zhou§

March 4, 2016

## Abstract

This article considers the sequential Monte Carlo (SMC) approximation of ratios of normalizing constants associated to posterior distributions which in principle rely on continuum models. Therefore, the Monte Carlo estimation error and the discrete approximation error must be balanced. A multilevel strategy is utilized to substantially reduce the cost to obtain a given error level in the approximation as compared to standard estimators. Two estimators are considered and relative variance bounds are given. The theoretical results are numerically illustrated for the example of identifying a parametrized permeability in an elliptic equation given point-wise observations of the pressure.

**Key words:** Multi-Level Monte Carlo, Sequential Monte Carlo, Bayesian Inverse Problems.

**AMS subject classification:** 82C80, 60K35.

## 1 Introduction

Over the past decades there has been an explosion of interest in accounting for uncertainty in the simulation of systems in science and engineering applications which are governed by continuum limiting systems such as partial differential equations (PDEs) [18, 23, 24]. The setting bears similarities to the case of continuous stochastic processes, which have enjoyed attention for much longer (e.g. [19]).

---

\*Center INRIA Bordeaux Sud-Ouest & Institut de Mathematiques de Bordeaux, Universite de Bordeaux I, 33405, FR

†Department of Statistics & Applied Probability, National University of Singapore, Singapore, 117546, SG

‡Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, 37934 TN, USA

§Department of Statistics & Applied Probability, National University of Singapore, Singapore, 117546, SG

Consider a sequence of probability measures  $\{\eta_l\}_{l \geq 0}$  on a common measurable space  $(E, \mathcal{E})$ ; assume that the probabilities have common dominating finite-measure  $du$ . In particular, for some known  $\kappa_l : E \rightarrow \mathbb{R}^+$ , let

$$\eta_l(du) = \frac{\kappa_l(u)}{Z_l} du \quad (1)$$

where the normalizing constant  $Z_l = \int_E \kappa_l(u) du$  may be unknown. The context of interest is when the sequence of densities is associated to an ‘accuracy’ parameter  $h_l$ , with  $h_l \rightarrow 0$  as  $l \rightarrow \infty$  with  $\infty > h_0 > h_1 > \dots > h_\infty = 0$ .

When estimating statistics  $\mathbb{E}_{\eta_\infty}[g(U)]$ , for  $g : E \rightarrow \mathbb{R}$ , in general one must approximate the limiting measure by  $\eta_L$  and perform statistical estimation with respect to this. For larger  $L$ , the approximation of the limit is better, and yet the simulations are more expensive and indeed the measure may also be supported on a subspace of the underlying space  $E$  whose dimension is larger.

Monte Carlo methods for statistical estimation are robust and scalable, although they are plagued by a “slow” convergence rate of  $\mathcal{O}(N^{-1/2})$  for approximations using  $N$  degrees of freedom. Attempts to circumvent this issue, for example using sophisticated deterministic high-dimensional approximation methods typically result in some manifestation of the “curse of dimensionality” [2], although recent work has indicated potential for the mitigation of such effect for suitably regular problems [6, 21].

The multilevel Monte Carlo framework [10, 11, 12] allows one to leverage in an optimal way the nested problems arising in this context, hence minimizing the necessary cost to obtain a given level of mean square error. In particular, the multilevel Monte Carlo method seeks to sample from  $\eta_0$  as well as a sequence of coupled pairs  $(\eta_0, \eta_1), \dots, (\eta_{L-1}, \eta_L)$  and using a collapsing sum representation of  $\mathbb{E}_{\eta_L}[g(U)]$ . Then using a suitable trade off of computational effort, one can reduce the amount of work, relative to i.i.d. sampling from  $\eta_L$  and using Monte Carlo integration, for a given amount of error. However, we are concerned with the scenario where such independent sampling is not possible, that is, either  $\eta_L$  or from the sequence of couples. As it is well-known, the use of importance sampling to then use the collapsing sum representation, is often not reasonable, in the sense that for importance proposals that can be sampled independently, the associated variance typically explodes exponentially in the dimension of the problem (e.g. [5]). As a result, there has been an extension of multilevel Monte Carlo methods in which the approximate target distribution can be sampled from directly, to more sophisticated Monte Carlo techniques for inference; however, this is still in its infancy. Important examples include the preliminary exploration of multilevel Markov chain Monte Carlo (MCMC) [13, 17], multilevel sequential Monte Carlo samplers [4], multilevel ensemble Kalman filter [14] and multilevel particle filters [16]. It should be noted that MCMC and SMC can perform at a polynomial cost in the dimension; see e.g. [3] and the references therein.

A significant challenge for inference problems is estimation of the normalizing constant  $Z_L$  or ratios thereof  $Z_l/Z_k$ ,  $L \geq l > k \geq 0$ . Such quantities are central to Bayesian model comparison and choice [15, 25]. In addition, obtaining

unbiased estimates (in the sense that the expectation is equal to the value, that is, potentially including discretization bias) are often central in pseudo-marginal algorithms (e.g. [1]). In general the calculation of these quantities are notoriously challenging (see for instance [26]) from a computational perspective.

In this article we extend the framework of [4] to consider the estimation of the ratio of normalizing constants. This is a framework which uses SMC. We consider both the ‘standard’ unbiased estimator ([8]) used in SMC, adapted to the multilevel setting and an estimator which follows the collapsing sum approach for multilevel methods. For the latter, we introduce a novel decomposition of the normalizing constant of a Feynman-Kac formula, which corresponds to  $Z_L/Z_0$ , which facilitates unbiased estimation. We consider new variance bounds for the estimator [8] and our new estimate and show that, in general, both approaches perform in a similar manner. For a given level of error, the cost is less than a Monte Carlo estimate that uses i.i.d. sampling from  $\eta_0$ , to estimate  $Z_L/Z_0$ ; we assume that the former is possible.

The paper is structured as follows. In Section 2 the setup will be given, along with a description of the multilevel algorithm and the new novel estimator for the normalizing constant. Section 3 contains the theoretical results, including the main theorems of the paper which allow the multilevel theory to carry through. Finally, section 4 presents the results of numerical experiments on an example Bayesian inverse problem. The proofs are housed in the appendix.

## 2 Estimation

### 2.1 Notations

Let  $(E, \mathcal{E})$  be a measurable space. The notation  $\mathcal{B}_b(E)$  denotes the class of bounded and measurable real-valued functions. The supremum norm is written as  $\|f\|_\infty = \sup_{u \in E} |f(u)|$  and  $\mathcal{P}(E)$  is the set of probability measures on  $(E, \mathcal{E})$ . We will consider non-negative operators  $K : E \times \mathcal{E} \rightarrow \mathbb{R}_+$  such that for each  $u \in E$  the mapping  $A \mapsto K(u, A)$  is a finite non-negative measure on  $\mathcal{E}$  and for each  $A \in \mathcal{E}$  the function  $u \mapsto K(u, A)$  is measurable; the kernel  $K$  is Markovian if  $K(u, dv)$  is a probability measure for every  $u \in E$ . For a finite measure  $\mu$  on  $(E, \mathcal{E})$ , and a real-valued, measurable  $f : E \rightarrow \mathbb{R}$ , we define the operations:

$$\mu K : A \mapsto \int K(u, A) \mu(du) ; \quad Kf : u \mapsto \int f(v) K(u, dv).$$

We also write  $\mu(f) = \int f(u) \mu(du)$ .

### 2.2 Algorithm

As described in the Introduction, the context of interest is when a sequence of densities  $\{\eta_l\}_{l \geq 0}$ , as in (1), are associated to an ‘accuracy’ parameter  $h_l$ , with  $h_l \rightarrow 0$  as  $l \rightarrow \infty$ , such that  $\infty > h_0 > h_1 \cdots > h_\infty = 0$ . In practice one cannot treat  $h_\infty = 0$  and so must consider these distributions with  $h_l > 0$ . The laws

with large  $h_l$  are easy to sample from with low computational cost, but are very different from  $\eta_\infty$ , whereas, those distributions with small  $h_l$  are hard to sample with relatively high computational cost, but are closer to  $\eta_\infty$ . Thus, we choose a maximum level  $L \geq 1$  and we will estimate

$$\mathbb{E}_{\eta_L}[g(U)] := \int_E g(u) \eta_L(du) .$$

By the standard telescoping identity used in MLMC, one has

$$\begin{aligned} \mathbb{E}_{\eta_L}[g(U)] &= \mathbb{E}_{\eta_0}[g(U)] + \sum_{l=1}^L \left\{ \mathbb{E}_{\eta_l}[g(U)] - \mathbb{E}_{\eta_{l-1}}[g(U)] \right\} \\ &= \mathbb{E}_{\eta_0}[g(U)] + \sum_{l=1}^L \mathbb{E}_{\eta_{l-1}} \left[ \left( \frac{\kappa_l(U) Z_{l-1}}{\kappa_{l-1}(U) Z_l} - 1 \right) g(U) \right] . \end{aligned} \quad (2)$$

Suppose now that one applies an SMC sampler [9] to obtain a collection of samples (particles) that sequentially approximate  $\eta_0, \eta_1, \dots, \eta_L$ . We consider the case when one initializes the population of particles by sampling i.i.d. from  $\eta_0$ , then at every step resamples and applies a MCMC Markov kernel to mutate the particles. We denote by  $(U_0^{1:N_0}, \dots, U_{L-1}^{1:N_{L-1}})$ , with  $+\infty > N_0 \geq N_1 \geq \dots \geq N_{L-1} \geq 1$ , the samples after mutation; one resamples  $U_l^{1:N_l}$  according to the weights  $G_l(U_l^i) = (\kappa_{l+1}/\kappa_l)(U_l^i)$ , for indices  $l \in \{0, \dots, L-1\}$ . We will denote by  $\{M_l\}_{1 \leq l \leq L-1}$  the sequence of MCMC kernels used at stages  $1, \dots, L-1$ , such that  $\eta_l M_l = \eta_{l+1}$ . For  $\varphi : E \rightarrow \mathbb{R}$ ,  $l \in \{1, \dots, L\}$ , we have the following estimator of  $\mathbb{E}_{\eta_{l-1}}[\varphi(U)]$ :

$$\eta_{l-1}^{N_{l-1}}(\varphi) = \frac{1}{N_{l-1}} \sum_{i=1}^{N_{l-1}} \varphi(U_{l-1}^i) .$$

We define

$$\eta_{l-1}^{N_{l-1}}(G_{l-1} M_l(du_l)) = \frac{1}{N_{l-1}} \sum_{i=1}^{N_{l-1}} G_{l-1}(U_{l-1}^i) M_l(U_{l-1}^i, du_l) .$$

The joint probability distribution for the SMC algorithm is

$$\prod_{i=1}^{N_0} \eta_0(du_0^i) \prod_{l=1}^{L-1} \prod_{i=1}^{N_l} \frac{\eta_{l-1}^{N_{l-1}}(G_{l-1} M_l(du_l^i))}{\eta_{l-1}^{N_{l-1}}(G_{l-1})} .$$

The algorithm is summarized in Figure 1. If one considers one more step in the above procedure, that would deliver samples  $\{U_L^i\}_{i=1}^{N_L}$ , a standard SMC sampler estimate of the quantity of interest in (2) is  $\eta_L^{N_L}(g)$ ; the earlier samples are discarded. Within a multi-level context, a consistent SMC estimate of (2) is

$$\hat{Y} = \eta_0^{N_0}(g) + \sum_{l=1}^L \left\{ \frac{\eta_{l-1}^{N_{l-1}}(g G_{l-1})}{\eta_{l-1}^{N_{l-1}}(G_{l-1})} - \eta_{l-1}^{N_{l-1}}(g) \right\} , \quad (3)$$

The motivation for such a procedure is that, as shown in [4], the amount of work, for a given level of error, relative to i.i.d. sampling from  $\eta_L$  is reduced. Thus the idea of using the approach is clear. However, as is well known in the literature (e.g. [9]) SMC samplers can also estimate ratios of normalizing constants as a by-product of the algorithm. We now consider this and also the amount of work to obtain a given level of error in this context.

- 
0. Sample  $U_0^1, \dots, U_0^{N_0}$  i.i.d. from  $\eta_0$  and compute  $G_0(u_0^i)$  for each sample  $i \in \{1, \dots, N_0\}$ : Set  $l = 0$ .
  1. Sample  $\check{U}_l^1, \dots, \check{U}_l^{N_{l+1}}$  with replacement from  $u_l^{1:N_l}$  with selection probabilities  $\{G_l(u_l^1)/\sum_{j=1}^{N_l} G_l(u_l^j), \dots, (G_l(u_l^{N_l})/\sum_{j=1}^{N_l} G_l(u_l^j))\}$ .
  2. Sample  $U_{l+1}^i | \check{u}_l^i$  from  $M_{l+1}(\check{u}_l^i, \cdot)$  and compute  $G_{l+1}(u_{l+1}^i)$  for each sample  $i \in \{1, \dots, N_{l+1}\}$ .
  3. Set  $l = l + 1$ . If  $l = L$  stop, otherwise return to the start of Step 1.
- 

Figure 1: The SMC algorithm.

## 2.3 Normalizing Constant

Define, for  $l \geq 0$

$$\gamma_l(du_l) = \int_{E^l} \left( \prod_{p=0}^{l-1} G_p(u_p) \right) \eta_0(du_l) \prod_{p=1}^l M_p(u_{p-1}, du_p).$$

In our context, it is well known that:

$$\gamma_l(1) = \frac{Z_l}{Z_0} = \prod_{p=0}^{l-1} \eta_p(G_p).$$

This suggests the estimator:

$$\gamma_l^{N_{0:l-1}}(1) = \prod_{p=0}^{l-1} \eta_p^{N_p}(G_p) \quad (4)$$

which is known to be unbiased ([8]). We consider the relative variance of this estimator in Section 3. However, at least on appearance it may not take advantage of the nature of the ML method. In addition, we show that the new estimator below, can potentially be leveraged to remove the discretization bias.

We propose the following procedure. It should be remarked that it holds in the particular context under study, but not for general Feynman-Kac models as

will be explained below. We have that for any  $l \geq 2$

$$\begin{aligned}
\gamma_l(1) &= \eta_0(G_0) + \sum_{p=2}^l \left( \gamma_p(1) - \gamma_{p-1}(1) \right) \\
&= \eta_0(G_0) + \sum_{p=2}^l \left( \gamma_{p-2}(G_{p-2}(M_{p-1}(G_{p-1}) - 1)) \right) \\
&= \eta_0(G_0) + \sum_{p=2}^l \left( \gamma_{p-2}(G_{p-2}(G_{p-1} - 1)) \right).
\end{aligned}$$

It is the final line that we will approximate with our MLSMC sampler. It is noted that the final line holds in the specific case of interest, but is not generally true for a given Feynman-Kac formula. The proposed approximation is

$$\tilde{\gamma}_l^{N_{0:l-2}}(1) = \eta_0^{N_0}(G_0) + \sum_{p=2}^l \left( \gamma_{p-2}^{N_{0:p-2}}(G_{p-2}(G_{p-1} - 1)) \right)$$

where for any  $g \in \mathcal{B}_b(E)$ ,  $p \geq 2$

$$\gamma_{p-2}^{N_{0:p-2}}(g) = \left( \prod_{k=0}^{p-3} \eta_k^{N_k}(G_k) \right) \eta_{p-2}^{N_{p-2}}(g).$$

Note that for  $l \geq 2$ , one has, almost surely,

$$\tilde{\gamma}_l^{N_{0:l-1}}(1) \neq \prod_{p=0}^{l-1} \eta_p^{N_p}(G_p).$$

Using [8] it clearly follows that

$$\gamma_l(1) = \mathbb{E}[\tilde{\gamma}_l^{N_{0:l-2}}(1)]$$

where  $\mathbb{E}$  is the expectation w.r.t. the law of the SMC algorithm; the estimator is unbiased.

## 2.4 Biased Estimator

Noting the estimator (3) another alternative estimator of  $\gamma_l(1)$  is

$$\prod_{p=0}^{l-1} \left( \eta_0^{N_0}(G) + \sum_{l=1}^p \left\{ \frac{\eta_{l-1}^{N_{l-1}}(G_p G_{l-1})}{\eta_{l-1}^{N_{l-1}}(G_{l-1})} - \eta_{l-1}^{N_{l-1}}(G_p) \right\} \right).$$

One can easily prove that this estimate is consistent, but biased, in the sense that

$$\mathbb{E} \left[ \prod_{p=0}^{l-1} \left( \eta_0^{N_0}(G) + \sum_{l=1}^p \left\{ \frac{\eta_{l-1}^{N_{l-1}}(G_p G_{l-1})}{\eta_{l-1}^{N_{l-1}}(G_{l-1})} - \eta_{l-1}^{N_{l-1}}(G_p) \right\} \right) \right] \neq \gamma_l(1).$$

However, the main reason why one may not want to consider its use is due to the cost of computing this estimate. If  $\sum_{p=0}^{l-1} N_p C_p$  is the ordinary cost of computing (4) ( $C_p$  is the cost per sample), then the cost of this estimator is  $\sum_{p=0}^{l-1} N_p \sum_{q=p}^{l-1} C_q$ . Such a procedure is undesirable in general and this is not investigated further.

## 2.5 Estimator with no Discretization Bias

Let  $M \in \{1, 2, \dots\}$  be a random variable that is independent of the MLSMC algorithm with  $\mathbb{P}_M(M \geq m) > 0 \forall m > 0$ . Suppose further that one can prove for  $N_0, N_1, \dots$  fixed that

$$\lim_{p \rightarrow \infty} \mathbb{E} \left[ \left( \gamma_{p-2}^{N_{0:p-2}} (G_{p-2} G_{p-1}) - \gamma_\infty(1) \right)^2 \right]^{1/2} = 0 \quad (5)$$

$$\lim_{p \rightarrow \infty} \mathbb{E} \left[ \left( \gamma_{p-2}^{N_{0:p-2}} (G_{p-2}) - \gamma_\infty(1) \right)^2 \right]^{1/2} = 0 \quad (6)$$

$$\sum_{p=2}^{\infty} \frac{1}{\mathbb{P}_M(M \geq p)} \mathbb{E} \left[ \left( \gamma_{p-2}^{N_{0:p-2}} (G_{p-2} (G_{p-1}) - \gamma_\infty(1)) \right)^2 \right] < +\infty \quad (7)$$

then one can use the estimator from [20] to obtain an unbiased estimator for  $\gamma_\infty(1)$ :

$$\frac{1}{\mathbb{P}_M(M \geq 1)} \eta_0^{N_0}(G_0) + \sum_{p=2}^M \frac{1}{\mathbb{P}_M(M \geq p)} \left( \gamma_{p-2}^{N_{0:p-2}} (G_{p-2} (G_{p-1} - 1)) \right).$$

Note that, even if one can prove (5)-(7), one must be prepared to spend an arbitrary amount of computational cost, which is not reasonable in the current context. Hence we do not consider this further here. We further remark that this particular approach is unlikely to work when estimating  $\mathbb{E}_{\eta_\infty}[g(U)]$  (as in (2)) as there is no unbiased property of the estimators (unbiased in the sense of expectations and not associated to the discretization).

## 3 Theory

### 3.1 Relative Variance Bounds

Throughout  $E$  is compact. We make the following assumptions:

(A1) There exist  $0 < \underline{C} < \overline{C} < +\infty$  such that

$$\begin{aligned} \sup_{0 \leq l < \infty} \sup_{u \in E} \kappa_l(u) &\leq \overline{C} ; \\ \inf_{0 \leq l < \infty} \inf_{u \in E} \kappa_l(u) &\geq \underline{C} . \end{aligned}$$

(A2) There exists a  $\rho \in (0, 1)$  such that for any  $l \geq 1$ ,  $(u, v) \in E^2$ ,  $A \in \mathcal{E}$ :

$$\int_A M_l(u, du') \geq \rho \int_A M_l(v, dv') .$$

These assumptions are almost identical to those in [4]. (A1) is different but equivalent to (A1) in [4]. The proofs of the following Theorems are in Appendices B and C respectively. It is remarked that there are other results in the spirit of Theorem 1 below, (see [7, 22]) but the bounds are not sharp enough for the purposes of this work.

**Theorem 1.** *Assume (A1-2). Then there exists a  $c, C < +\infty$  such that for any  $L \geq 2$ ,  $N_0 \geq N_1 \geq \dots \geq N_{L-1} > cL$ ,*

$$\mathbb{E} \left[ \left( \frac{\gamma_L^{N_0:L-1}(1)}{\gamma_L(1)} - 1 \right)^2 \right] \leq C \sum_{p=0}^{L-1} \frac{1}{N_p} \left( \left( \sum_{q=p}^{L-1} \left\| \frac{G_q}{\eta_q(G_q)} - 1 \right\|_\infty \right)^2 + \left\| \frac{G_p}{\eta_p(G_p)} - 1 \right\|_\infty \frac{(p+1)}{N_p} \right).$$

**Theorem 2.** *Assume (A1-2). Then there exists a  $c, C < +\infty$  such that for any  $L \geq 2$ ,  $N_0 \geq N_1 \geq \dots \geq N_{L-2} > c(L-1)$ ,*

$$\mathbb{E} \left[ \left( \frac{\tilde{\gamma}_L^{N_0:L-2}(1)}{\gamma_L(1)} - 1 \right)^2 \right] \leq C \left( \frac{1}{N_0} + \sum_{p=2}^L \frac{(p-1)}{N_{p-2}} \|G_{p-1} - 1\|_\infty^2 + \sum_{p=2}^L \sum_{q=2}^{p-1} \frac{(q-1)}{N_{q-2}} \|G_{p-1} - 1\|_\infty \|G_{q-1} - 1\|_\infty \right).$$

### 3.2 Cost Analysis

In order to investigate the cost for a given level of error, we introduce the following assumption.

(A3) (i) There exist  $\alpha, \zeta > 0$ , and a  $C > 0$  such that for all  $p > 0$

$$\begin{cases} \left| \frac{\gamma_p(1)}{\gamma_\infty(1)} - 1 \right| & \leq Ch_p^\alpha; \\ C(G_{p-1}) & \leq Ch_p^{-\zeta}, \end{cases} \quad (8)$$

where  $C(G_{p-1})$  denotes the cost to evaluate  $G_{p-1}$ .

(ii) There exist a  $\beta > 0$  and a  $C > 0$  such that for all  $p > 0$

$$\left\| \frac{G_{p-1}}{\eta_{p-1}(G_{p-1})} - 1 \right\|_\infty^2 \leq Ch_p^\beta.$$

(iii) There exist a  $\beta > 0$  and a  $C > 0$  such that for all  $p > 0$

$$\|G_{p-1} - 1\|_\infty^2 \leq Ch_p^\beta.$$



**Corollary 3.1.** Assume (A1,2,3(i)(ii)) and  $2\alpha \geq \max\{\beta, \zeta\}$ . Then for any  $\varepsilon > 0$ , there exist  $L, \{N_l\}_{l=0}^L$  and  $C > 0$  such that

$$\frac{1}{\gamma_\infty(1)^2} \mathbb{E} \left[ \left( \gamma_L^{N_{0:L-2}}(1) - \gamma_\infty(1) \right)^2 \right] \leq C\varepsilon^2, \quad (9)$$

for the following cost

$$\text{COST} \leq C \begin{cases} \varepsilon^{-2} |\log(\varepsilon)|, & \text{if } \beta > \zeta, \\ \varepsilon^{-2} |\log(\varepsilon)|^3, & \text{if } \beta = \zeta, \\ \varepsilon^{-(2+\frac{\zeta-\beta}{\alpha})} |\log(\varepsilon)|, & \text{if } \beta < \zeta. \end{cases} \quad (10)$$

**Corollary 3.2.** Assume (A1,2,3(i)(iii)) and  $2\alpha \geq \max\{\beta, \zeta\}$ . Then for any  $\varepsilon > 0$ , there exist  $L, \{N_l\}_{l=0}^L$  and  $C > 0$  such that

$$\frac{1}{\gamma_\infty(1)^2} \mathbb{E} \left[ \left( \tilde{\gamma}_L^{N_{0:L-2}}(1) - \gamma_\infty(1) \right)^2 \right] \leq C\varepsilon^2, \quad (11)$$

for the following cost

$$\text{COST} \leq C \begin{cases} \varepsilon^{-2} |\log(\varepsilon)|, & \text{if } \beta > \zeta, \\ \varepsilon^{-2} |\log(\varepsilon)|^3, & \text{if } \beta = \zeta, \\ \varepsilon^{-(2+\frac{\zeta-\beta}{\alpha})} |\log(\varepsilon)|, & \text{if } \beta < \zeta. \end{cases} \quad (12)$$

We give the proof for Corollary 3.2 only. The proof of Corollary 3.1 is almost identical. The only difference is treating the term in the relative variance of

$$\sum_{p=0}^{L-1} \left\| \frac{G_p}{\eta_p(G_p)} - 1 \right\|_\infty \frac{(p+1)}{N_p^2}$$

which is smaller than  $\mathcal{O}(\varepsilon^2)$ , under our assumptions.

*Proof of Corollary 3.2.* The MSE can be bounded by

$$\begin{aligned} & \frac{1}{\gamma_\infty(1)^2} \mathbb{E} \left[ \left( \tilde{\gamma}_L^{N_{0:L-2}}(1) - \gamma_\infty(1) \right)^2 \right] \leq \\ & \left( \frac{\gamma_L(1)}{\gamma_\infty(1)} \right)^2 \mathbb{E} \left[ \left( \frac{\tilde{\gamma}_L^{N_{0:L-2}}(1)}{\gamma_L(1)} - 1 \right)^2 \right] + \left| \left( \frac{\gamma_L(1)}{\gamma_\infty(1)} - 1 \right) \right|^2. \end{aligned}$$

Following from (A3(i)), the second term requires that  $h_L^\alpha \approx \varepsilon$ , and assuming  $h_L = M^{-L}$  for some  $M \geq 2$ , this translates to  $L \approx \log \varepsilon$ . Notice that it also follows that  $\left( \frac{\gamma_L(1)}{\gamma_\infty(1)} \right)^2 = \mathcal{O}(1)$ . Now, defining  $V_p = \|G_{p-1} - 1\|_\infty^2$ , Theorem 2 provides the following bound for the first term

$$\mathbb{E} \left[ \left( \frac{\tilde{\gamma}_L^{N_{0:L-2}}(1)}{\gamma_L(1)} - 1 \right)^2 \right] \leq V := C \left( \frac{1}{N_0} + L \sum_{p=1}^{L-1} \frac{V_p}{N_{p-1}} \right).$$

To see this observe that

$$\sum_{p=1}^{L-1} \sum_{q=1}^p \frac{q}{N_{q-1}} V_p^{1/2} V_q^{1/2} = \sum_{p=1}^{L-1} \frac{p}{N_{p-1}} V_p^{1/2} \sum_{q=p}^{L-1} V_q^{1/2} \leq CL \sum_{p=1}^{L-1} \frac{V_p}{N_{p-1}}.$$

Optimizing the cost, given that the variance is  $\mathcal{O}(\varepsilon^2)$ , dictates that  $N_l \propto \sqrt{LV_l/C_l} \approx L^{1/2} h_l^{(\beta+\zeta)/2}$ . The constraint then requires that  $N_l \propto L\varepsilon^{-2} K_L h_l^{(\beta+\zeta)/2}$ , where  $K_L = \sum_{l=1}^{L-1} h_l^{(\beta-\zeta)/2}$ . By assumption  $\max\{\beta, \zeta\} \leq 2\alpha$ , so  $(\beta+\zeta)/2\alpha \leq 2$  and the requirement for all the  $N_l$  in Theorem 2 is guaranteed (as long as the proportionality constant is greater than 1). Therefore, the MSE is controlled by  $\mathcal{O}(\varepsilon^2)$  with a cost given by

$$\sum_{l=0}^L N_l C_l \approx L\varepsilon^{-2} K_L^2,$$

and the result follows.  $\square$

**Remark 1.** *If one were able to perform i.i.d. sampling from  $\eta_0$  (denote the samples  $u^1, \dots, u^N$ ), with estimator*

$$\frac{1}{N} \sum_{i=1}^N \frac{\gamma_L(u^i)}{\gamma_0(u^i)}$$

*for  $Z_L/Z_0$  a computational effort proportional to  $Nh_L^{-\zeta}$  is used, with  $N$  the number of simulated samples. To make the overall error (bias squared plus variance) of using i.i.d. sampling  $\mathcal{O}(\varepsilon^2)$  then one must take  $N \propto \mathcal{O}(\varepsilon^{-2})$ , as the variance of the MC estimate is  $\mathcal{O}(N^{-1})$ , independently of  $L$ . This is a computational cost of  $\mathcal{O}(\varepsilon^{-2} h_L^{-\zeta})$  is used which is far worse than MLSMC samplers in most cases of practical interest.*

## 4 Numerical Example

### 4.1 Setup

The performance of the proposed estimator will be demonstrated by a Bayesian inverse problem example. The same example was also used in [4], which introduced the MLSMC algorithm.

Introduce the Gelfand triple  $V := H^1(D) \subset L^2(D) \subset H^{-1}(D) =: V^*$ , where the domain  $D$  will be understood. Let  $D \subset \mathbb{R}^d$  with  $\partial D \in C^1$  convex. For  $f \in V^*$ , consider the following PDE on  $D$ :

$$-\nabla \cdot (\hat{u} \nabla p) = f, \quad \text{on } D, \tag{13}$$

$$p = 0, \quad \text{on } \partial D, \tag{14}$$

where

$$\hat{u}(x) = \bar{u}(x) + \sum_{k=1}^K u_k \sigma_k \phi_k(x). \quad (15)$$

Define  $u = \{u_k\}_{k=1}^K$ , with  $u_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{U}[-1, 1]$  (the uniform distribution on  $[-1, 1]$ ). Assume that  $\bar{u}, \phi_k \in C^\infty$  for all  $k$  and  $\|\phi_k\| = 1$ . In particular  $\{\sigma_k\}_{k=1}^K$  decay with  $k$ . In addition, the following property shall hold:

$$\inf_x \hat{u}(x) \geq \inf_x \bar{u}(x) - \sum_{k=1}^K \sigma_k \geq u_* > 0 \quad (16)$$

so that the operator on the left-hand side of Equation (13) is uniformly elliptic. Let  $p(\cdot; u)$  denote the weak solution of Equation (13) for parameter  $u$ . Define the following vector-valued function

$$\mathcal{G}(p) = [g_1(p), \dots, g_M(p)]^\top,$$

where  $g_m$  are elements of the dual space  $V^*$  for  $m = 1, \dots, M$ . It is assumed that the data take the form

$$y = \mathcal{G}(p) + \xi, \quad \xi \sim \mathcal{N}(0, \Xi), \quad (17)$$

where  $\mathcal{N}(0, \Xi)$  denotes the Normal distribution with zero mean and covariance  $\Xi$ .

The specific setting of the simulations are as the following. Let  $D = [0, 1]$  and  $f(x) = 100x$ . Set  $K = 50$ ,  $\bar{u}(x) = 0.15 = \text{const.}$ ,  $\sigma_k = (2/5)4^{-k} \phi_k(x) = \sin(k\pi x)$  if  $k$  is odd and  $\phi_k(x) = \cos(k\pi x)$  if  $k$  is even. The forward problem at resolution level  $l$  is solved using a finite element method with piecewise linear shape functions on a uniform mesh of with  $h_l = 2^{-(l+k)}$ , for some starting  $k \geq 1$  (so that there are at least two grid-blocks in the coarsest,  $l = 0$ , case). Thus, on level  $l$  the finite element basis functions are  $\{\psi_i^l\}_{i=1}^{2^{l+k}-1}$  defined as (for  $x_i = i \cdot 2^{-(l+k)}$ ):

$$\psi_i^l(x) = \begin{cases} (1/h_l)[x - (x_i - h_l)] & \text{if } x \in [x_i - h_l, x_i], \\ (1/h_l)[(x_i + h_l) - x] & \text{if } x \in [x_i, x_i + h_l]. \end{cases}$$

The function of interest  $g$  is taken as the solution of the forward problem at the midpoint of the domain, that is  $g(u) = p(0.5; u)$ . The observation operator is  $\mathcal{G}(u) = [p(0.25; u), p(0.75; u)]^\top$ , and the observational noise covariance is taken to be  $\Xi = 0.25^2 I$ .

Detailed error rates analysis of this example can be found in [4]. In particular, when the purpose of the study was to estimate  $\eta_L(g)$ , the variance rate was  $\beta = 4$  empirically. Later we will show that for estimating the normalizing constant, the variance rate is very similar.

## 4.2 Verification of Assumptions

Assumptions (A1) and (A3(i)(iii)) (for  $|\frac{\gamma_p(1)}{\gamma_\infty(1)} - 1|$ ), with  $\beta = 2\alpha = 2$ , follow from Proposition 4.1 of [4]. For (A3(ii)) this follows directly from proving (A3(iii)). It is natural to model the cost at level  $p$  by a power of the number degrees of freedom, which is in turn related to  $h_p^{-1}$ , verifying (A3(i)) (for  $C(G_{p-1})$ ). The stiffness matrix of the finite element method is tridiagonal and thus the system can be solved with cost  $\mathcal{O}(2^{l+k})$ , corresponding to a computational cost rate of  $\zeta = 1$ . Assumption (A2) is verified for Gibbs sampler in section 4.2 of [4].

## 4.3 Experiments

We begin by using the theoretical rates  $\beta = 2\alpha = 2$  to estimate the MSE and hence the cost ratio. Three cases are considered:

- A standard SMC algorithm, with the estimator  $\gamma_l^{N_{0:l-1}}(1)$ .
- MLSMC sampler for  $\gamma_l^{N_{0:l-1}}(1)$ .
- MLSMC sampler for  $\tilde{\gamma}_l^{N_{0:l-2}}(1)$ .

The cost vs. MSE is plotted in Figure 2. The cost rates are  $-1.271$ ,  $-0.967$ , and  $-1.038$  for the SMC, MLSMC with the standard estimator, and MLSMC with the new estimator, respectively. It is clear that the MLSMC algorithm with both estimators provides superior performance when compared to the standard SMC algorithm. It is interesting that for the given MLSMC ensemble, the performance of the new estimator is comparable to that of the standard estimator, as proven in Corollaries 3.1 and 3.2. It shall be noted that in practice, given the same samples  $(U_0^{1:N_0}, \dots, U_{L-1}^{1:N_{L-1}})$ , the new estimator is capable of estimating  $\gamma_{L+1}(1)$  while the standard one can only estimate the  $\gamma_L(1)$ , which has a higher bias.

The variance rate  $\beta$  can also be estimated empirically by consider the variance of  $\eta_l(G_l)$ . The quantity, multiplied by the sample size, as a proxy of  $V_l$  is plotted in Figure 3. The estimated empirical rate is  $\beta = 4.148$ . This is consistent with the rate estimates in [4].

## Acknowledgements

KJHL was supported in part by DARPA FORMULATE and in part by ORNL LDRD Strategic Hire. AJ & YZ were supported by Ministry of Education AcRF tier 2 grant, R-155-000-161-112.

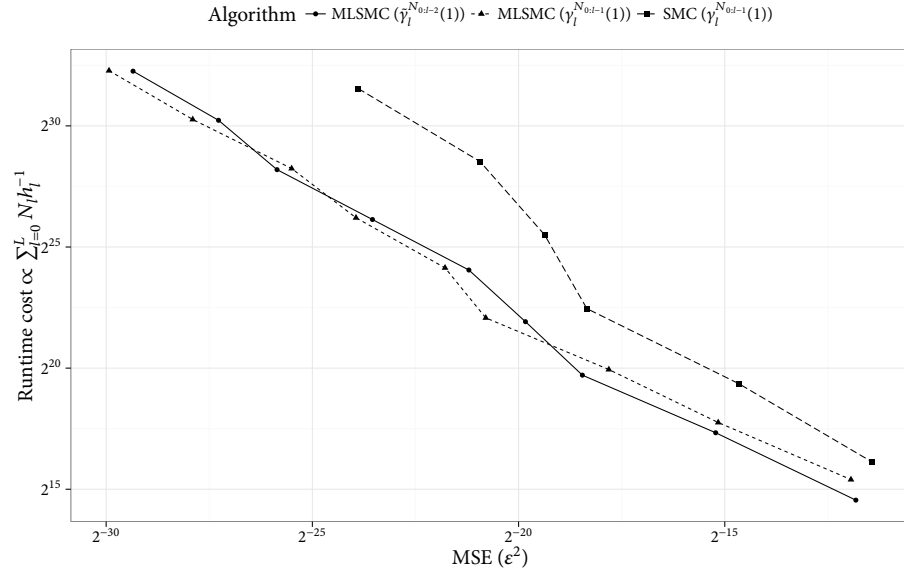


Figure 2: Computational cost against mean squared error

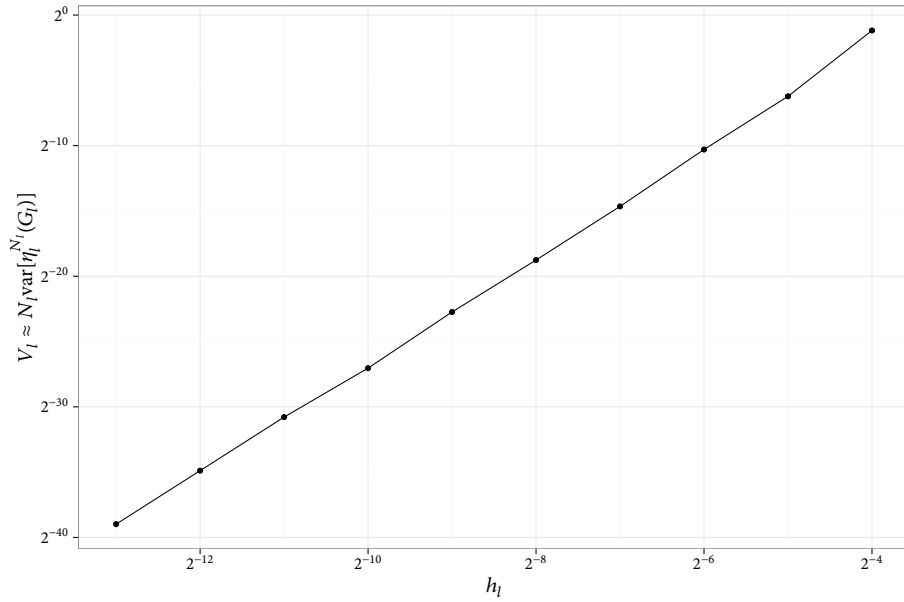


Figure 3: Variance rate estimate

## A Notations

We give a collection of definitions which are used in the appendices. Let  $n \geq 0$ ,  $F \in \mathcal{B}_b(E \times E)$  and define

$$(\gamma_n^{N_{0:n}})^{\otimes 2}(F) = \left( \prod_{p=0}^{n-1} \eta_p^{N_p}(G_p) \right)^2 (\eta_n^{N_n})^{\otimes 2}(F)$$

where for a finite (possibly signed) measure on  $E$ ,  $\mu, \mu^{\otimes 2}(d(u_1, u_2)) = \mu(du_1)\mu(du_2)$ . We recall the semi-group for  $p \leq n$  (for  $p = n$  it is the identity operator):

$$Q_{p,n}(x_p, dx_n) = \int_{E^{n-p-1}} Q_{p+1}(x_p, dx_{p+1}) \dots Q_n(x_{n-1}, dx_n)$$

where for  $n \geq 1$ ,  $Q_n(x, dy) = G_{n-1}(x)M_n(x, dy)$ . We also define the coalescent operator for  $F \in \mathcal{B}_b(E \times E)$ ,  $(x, y) \in E \times E$ :

$$C(F)(x, y) = F(x, x).$$

Then for  $0 \leq s \leq (n+1)$ ,  $0 \leq i_1 < \dots < i_n \leq n$ ,  $F \in \mathcal{B}_b(E \times E)$

$$\Gamma_n^{i_1:i_s}(F) = \gamma_{i_1}^{\otimes 2} C Q_{i_1, i_2}^{\otimes 2} C Q_{i_2, i_3}^{\otimes 2} \dots C Q_{i_s, n}^{\otimes 2}(F)$$

and

$$\bar{\Gamma}_n^{i_1:i_s}(F) = \frac{1}{\gamma_n(1)^2} \Gamma_n^{i_1:i_s}(F).$$

The conventions, for  $s = 0$ ,  $\Gamma_n^\emptyset(F) = \gamma_n^{\otimes 2}(F)$  and  $\bar{\Gamma}_n^\emptyset(F) = \eta_n^{\otimes 2}(F)$  are adopted. Recall the selection-mutation operator for any  $\mu \in \mathcal{P}(E)$ ,  $n \geq 1$

$$\Phi_n(\mu)(dx) = \frac{\mu(G_{n-1}M_n(\cdot, dx))}{\mu(G_{n-1})}.$$

$\mathcal{F}_n^{N_{0:n}}$  denotes the natural filtration generated by the particle system up-to time  $n$ . For  $f_1, f_2 \in \mathcal{B}_b(E)$  we write the tensor product of functions for every  $(x, y) \in E \times E$ :

$$f_1 \otimes f_2(x, y) = f_1(x)f_2(y).$$

## B Proofs for Theorem 1

**Lemma 1.** *Assume (A1-2). Then there exist a  $C < +\infty$  such that for any  $0 \leq p \leq n$ ,  $x \in E$ :*

$$\left| \frac{Q_{p,n}(1)(x)}{\prod_{q=p}^{n-1} \eta_q(G_q)} - 1 \right| \leq C \sum_{q=p}^{n-1} \left\| \frac{G_q}{\eta_q(G_q)} - 1 \right\|_\infty$$

*Proof.* We fix  $n, p$  and note that the case  $p = n$  is trivial, so we suppose  $p < n$ . We prove the result by induction. We consider  $p = n - 1$  and thus

$$\frac{Q_{p,n}(1)(x)}{\prod_{q=p}^{n-1} \eta_q(G_q)} - 1 = \frac{G_{n-1}(x)}{\eta_{n-1}(G_{n-1})} - 1$$

so the initialization is proved. Suppose the result holds at rank  $p$  and consider the case  $p - 1$ . We have

$$\begin{aligned} & \frac{Q_{p-1,n}(1)(x)}{\prod_{q=p-1}^{n-1} \eta_q(G_q)} - 1 = \\ & \left( \frac{G_{p-1}(x)}{\eta_{p-1}(G_{p-1})} - 1 \right) M_p \left( \frac{Q_{p,n}(1)(x)}{\prod_{q=p}^{n-1} \eta_q(G_q)} \right) (x) + M_p \left( \frac{Q_{p,n}(1)(x)}{\prod_{q=p}^{n-1} \eta_q(G_q)} - 1 \right) (x). \end{aligned}$$

By [7, Lemma 4.1]

$$\frac{Q_{p,n}(1)(x)}{\prod_{q=p}^{n-1} \eta_q(G_q)} \leq C \quad (18)$$

where  $C$  does not depend upon  $p, n$ . Thus, by applying the induction hypothesis and the above result it follows that:

$$\left| \frac{Q_{p-1,n}(1)(x)}{\prod_{q=p-1}^{n-1} \eta_q(G_q)} - 1 \right| \leq C \sum_{q=p-1}^{n-1} \left\| \frac{G_q}{\eta_q(G_q)} - 1 \right\|_{\infty}$$

and hence the proof is completed.  $\square$

The result below follows one in [7].

**Proposition 1.** *Assume (A1-2). Then there exists a  $C < +\infty$  such that for any  $n \geq 0$ ,  $F \in \mathcal{B}_b(E \times E)$  and  $N_0 \geq \dots \geq N_n > c(n+1)$*

$$\left| \mathbb{E} \left[ \frac{(\gamma_n^{N_{0:n}})^{\otimes 2}(F)}{\gamma_n(1)^2} \right] - \eta_n^{\otimes 2}(F) \right| \leq 8c \|F\|_{\infty} \sum_{p=0}^n \frac{1}{N_p}.$$

*Proof.* The case with  $F$  constant essentially follows from the proofs of [7]. The only difference is the fact that we have a decreasing number of samples; this does not change the calculations of that paper, so the case of  $F$  constant is in [7]. If  $F$  is a non-constant function, one has, from the equation above Proposition 3.4 (page 638) of [7]:

$$\begin{aligned} & \left| \mathbb{E} \left[ \frac{(\gamma_n^{N_{0:n}})^{\otimes 2}(F)}{\gamma_n(1)^2} \right] - \eta_n^{\otimes 2}(F) \right| = \\ & \left| \sum_{s=1}^{n+1} \sum_{0 \leq i_1 < \dots < i_s \leq n} \left( \prod_{k=1}^s \frac{1}{N_{i_k}} \right) \left( \prod_{k \notin \{i_1, \dots, i_s\}} \left( 1 - \frac{1}{N_k} \right) \right) \bar{\Gamma}_n^{i_1:i_s}(F - \eta_n^{\otimes 2}(F)) \right|. \end{aligned}$$

Following the proof of Theorem 5.1 of [7] and noting that one can allow the function in that paper to be negative, it follows that

$$|\bar{\Gamma}_n^{i_1:i_s}(F - \eta_n^{\otimes 2}(F))| \leq \|F - \eta_n^{\otimes 2}(F)\|_\infty \left(\rho \frac{\bar{C}}{\underline{C}}\right)^s \leq 2\|F\|_\infty \left(\rho \frac{\bar{C}}{\underline{C}}\right)^s.$$

Thus one has

$$\left| \mathbb{E} \left[ \frac{(\gamma_n^{N_{0:n}})^{\otimes 2}(F)}{\gamma_n(1)^2} \right] - \eta_n^{\otimes 2}(F) \right| \leq 2\|F\|_\infty \sum_{s=1}^{n+1} \sum_{0 \leq i_1 < \dots < i_s \leq n} \left( \prod_{k=1}^s \frac{1}{N_{i_k}} \right) \left( \rho \frac{\bar{C}}{\underline{C}} \right)^s.$$

Note that

$$\sum_{s=1}^{n+1} \sum_{0 \leq i_1 < \dots < i_s \leq n} \left( \prod_{k=1}^s \frac{1}{N_{i_k}} \right) \left( \rho \frac{\bar{C}}{\underline{C}} \right)^s = \prod_{s=0}^n \left( 1 + \rho \frac{\bar{C}}{\underline{C}} \frac{1}{N_s} \right) - 1,$$

and for  $N_0 > C(n+1), \dots, N_n > C(n+1)$

$$\prod_{s=0}^n \left( 1 + \rho \frac{\bar{C}}{\underline{C}} \frac{1}{N_s} \right) - 1 \leq 2\rho \frac{\bar{C}}{\underline{C}} \sum_{p=0}^n \frac{1}{N_p},$$

(see for instance the proofs of Theorem 5.1 and Corollary 5.2 of [7]). It follows that

$$\left| \mathbb{E} \left[ \frac{(\gamma_n^{N_{0:n}})^{\otimes 2}(F)}{\gamma_n(1)^2} \right] - \eta_n^{\otimes 2}(F) \right| \leq 8C\|F\|_\infty \sum_{p=0}^n \frac{1}{N_p},$$

with  $C = \rho \frac{\bar{C}}{\underline{C}}$ ; the proof is concluded.  $\square$

*Proof of Theorem 1.* Throughout the proof  $C < +\infty$  is a constant whose value may change from line-to-line. It will not depend on the level index. By [22, Proposition 2.3]

$$\mathbb{E} \left[ \left( \frac{\gamma_L^{N_{0:L-1}}(1)}{\gamma_L(1)} - 1 \right)^2 \right] = \sum_{p=0}^{L-1} \frac{1}{N_p} \mathbb{E}[T_{p,L}^{N_{0:p}}] \quad (19)$$

where

$$T_{p,L}^{N_{0:p}} = \left( \frac{\gamma_p^{N_{0:p-1}}(1)}{\gamma_p(1)} \right)^2 \left( \eta_p^{N_p}(h_{p,L}^2) - \eta_p^{N_p}(h_{p,L})^2 + \eta_p^{N_p}(h_{p,L}) \eta_p^{N_p} \left( \frac{G_p}{\eta_p(G_p)} - 1 \right) \right)$$

and we use the short-hand for  $0 \leq p \leq n$ ,  $x \in E$ :

$$h_{p,n}(x) = \frac{Q_{p,n}(1)(x)}{\prod_{q=p}^{n-1} \eta_q(G_q)}.$$

Now, one has almost surely that

$$T_{p,L}^{N_{0:p}} = \left( \frac{\gamma_p^{N_{0:p-1}}(1)}{\gamma_p(1)} \right)^2 \times$$



$$\left( \eta_p^{N_p}([h_{p,L} - 1]^2) - \eta_p^{N_p}(h_{p,L} - 1)^2 + \eta_p^{N_p}(h_{p,L})\eta_p^{N_p}\left(\frac{G_p}{\eta_p(G_p)} - 1\right) \right).$$

As

$$\begin{aligned} \mathbb{E}[T_{p,L}^{N_{0:p}}] &= \mathbb{E} \left[ \left( \frac{\gamma_p^{N_{0:p-1}}(1)}{\gamma_p(1)} \right)^2 \left( \eta_p^{N_p}([h_{p,L} - 1]^2) - \eta_p^{N_p}(h_{p,L} - 1)^2 \right) \right] + \\ &\quad \mathbb{E} \left[ \left( \frac{\gamma_p^{N_{0:p-1}}(1)}{\gamma_p(1)} \right)^2 \eta_p^{N_p}(h_{p,L})\eta_p^{N_p}\left(\frac{G_p}{\eta_p(G_p)} - 1\right) \right] \end{aligned} \quad (20)$$

we will consider controlling the two terms on the R.H.S. of (20) separately.

**First term on the R.H.S. of (20).**

We have, almost surely that

$$\begin{aligned} \eta_p^{N_p}([h_{p,L} - 1]^2) - \eta_p^{N_p}(h_{p,L} - 1)^2 &\leq C \|h_{p,L} - 1\|_\infty^2 \\ &\leq C \left( \sum_{q=p}^{L-1} \left\| \frac{G_q}{\eta_p(G_q)} - 1 \right\|_\infty \right)^2 \end{aligned}$$

where we have applied Lemma 1 to go to the second line. Then by Proposition 1 as  $N_0 > cL, \dots, N_{L-1} > cL$

$$\mathbb{E} \left[ \left( \frac{\gamma_p^{N_{0:p-1}}(1)}{\gamma_p(1)} \right)^2 \right] \leq C.$$

So we have shown that

$$\begin{aligned} \mathbb{E} \left[ \left( \frac{\gamma_p^{N_{0:p-1}}(1)}{\gamma_p(1)} \right)^2 \left( \eta_p^{N_p}([h_{p,L} - 1]^2) - \eta_p^{N_p}(h_{p,L} - 1)^2 \right) \right] &\leq \\ C \left( \sum_{q=p}^{L-1} \left\| \frac{G_q}{\eta_p(G_q)} - 1 \right\|_\infty \right)^2. \end{aligned} \quad (21)$$

**Second term on the R.H.S. of (20).**

We have almost surely that

$$\begin{aligned} \left( \frac{\gamma_p^{N_{0:p-1}}(1)}{\gamma_p(1)} \right)^2 \eta_p^{N_p}(h_{p,L})\eta_p^{N_p}\left(\frac{G_p}{\eta_p(G_p)} - 1\right) &= \\ \frac{1}{\gamma_p(1)^2} (\gamma_p^{N_{0:p}})^{\otimes 2} \left( h_{p,L} \otimes \left( \frac{G_p}{\eta_p(G_p)} - 1 \right) \right) \end{aligned}$$

and note that,

$$\eta_p^{\otimes 2} \left( h_{p,L} \otimes \left( \frac{G_p}{\eta_p(G_p)} - 1 \right) \right) = 0.$$

So by Proposition 1 as  $N_0 > cL, \dots, N_{L-1} > cL$  and (18)

$$\left| \mathbb{E} \left[ \left( \frac{\gamma_p^{N_{0:p-1}}(1)}{\gamma_p(1)} \right)^2 \eta_p^{N_p}(h_{p,L}) \eta_p^{N_p} \left( \frac{G_p}{\eta_p(G_p)} - 1 \right) \right] \right| \leq C \left\| \frac{G_p}{\eta_p(G_p)} - 1 \right\|_\infty \frac{(p+1)}{N_p}. \quad (22)$$

Combining (19) with (20) and after applying the triangular inequality, the bounds (21) and (22) complete the proof.  $\square$

## C Proofs for Theorem 2

Some of the proofs in this Section will use Proposition 1 in Appendix B.

**Lemma 2.** *Let  $n \geq 1$  and  $f_1, f_2 \in \mathcal{B}_b(E)$  then*

$$\mathbb{E} \left[ [\gamma_n^{N_{0:n}} - \gamma_n](f_1) [\gamma_n^{N_{0:n}} - \gamma_n](f_2) \right] = \mathbb{E} [(\gamma_n^{N_{0:n}})^{\otimes 2}(f_1 \otimes f_2)] - \gamma_n(1)^2 \eta_n^{\otimes 2}(f_1 \otimes f_2).$$

*Proof.* We have

$$\begin{aligned} & \mathbb{E} \left[ [\gamma_n^{N_{0:n}} - \gamma_n](f_1) [\gamma_n^{N_{0:n}} - \gamma_n](f_2) \right] = \\ & \mathbb{E} [\gamma_n^{N_{0:n}}(f_1) \gamma_n^{N_{0:n}}(f_2)] - \gamma_n(f_2) \mathbb{E} [\gamma_n^{N_{0:n}}(f_1)] - \gamma_n(f_1) \mathbb{E} [\gamma_n^{N_{0:n}}(f_2)] + \gamma_n(f_1) \gamma_n(f_2) = \\ & \mathbb{E} [\gamma_n^{N_{0:n}}(f_1) \gamma_n^{N_{0:n}}(f_2)] - \gamma_n(f_2) \gamma_n(f_1) - \gamma_n(f_1) \gamma_n(f_2) + \gamma_n(f_1) \gamma_n(f_2) \end{aligned}$$

where the unbiased property of the normalizing constant has been used to go to the last line. Then it follows that

$$\begin{aligned} & \mathbb{E} [\gamma_n^{N_{0:n}}(f_1) \gamma_n^{N_{0:n}}(f_2)] - \gamma_n(f_2) \gamma_n(f_1) - \gamma_n(f_1) \gamma_n(f_2) + \gamma_n(f_1) \gamma_n(f_2) = \\ & \mathbb{E} [(\gamma_n^{N_{0:n}})^{\otimes 2}(f_1 \otimes f_2)] - \gamma_n(1)^2 \eta_n^{\otimes 2}(f_1 \otimes f_2) \end{aligned}$$

which concludes the proof.  $\square$

**Lemma 3.** *Assume (A1-2). Then there exists a  $C < +\infty$  such that for any  $2 \leq q < p$ ,  $N_0 \geq N_1 \geq \dots \geq N_{q-2} > C(q-1)$ :*

$$\begin{aligned} & |\mathbb{E} [[\gamma_{p-2}^{N_{0:p-2}} - \gamma_{p-2}](G_{p-2}(G_{p-1} - 1)) [\gamma_{q-2}^{N_{0:q-2}} - \gamma_{q-2}](G_{q-2}(G_{q-1} - 1))]| \leq \\ & \frac{c(q-1)\gamma_{q-2}(1)^2}{N_{q-2}} \left\| G_{q-2}(G_{q-1} - 1) Q_{q-2,p-2}(G_{p-2}(G_{p-1} - 1)) \right\|_\infty. \end{aligned}$$

*Proof.* From [8, Proposition 7.4.1] we have

$$\begin{aligned} & \mathbb{E} [[\gamma_{p-2}^{N_{0:p-2}} - \gamma_{p-2}](G_{p-2}(G_{p-1} - 1)) [\gamma_{q-2}^{N_{0:q-2}} - \gamma_{q-2}](G_{q-2}(G_{q-1} - 1))] = \\ & \sum_{s_1=0}^{p-2} \sum_{s_2=0}^{q-2} \mathbb{E} \left[ \gamma_{s_1}^{N_{0:s_1-1}}(1) [\eta_{s_1}^{N_{s_1}} - \Phi_{s_1}(\eta_{s_1-1}^{N_{s_1-1}})] (Q_{s_1,p-2}(\overline{G}_p)) \times \right. \end{aligned}$$

$$\gamma_{s_2}^{N_{0:s_2-1}}(1)[\eta_{s_2}^{N_{s_2}} - \Phi_{s_2}(\eta_{s_2-1}^{N_{s_2-1}})](Q_{s_2,q-2}(\overline{G}_q))\Big]$$

where we have used the shorthand  $\overline{G}_s = G_{s-2}(G_{s-1} - 1)$  for any  $s \geq 2$ . For any  $s \geq 0$ ,  $f \in \mathcal{B}_b(E)$

$$\mathbb{E}[\gamma_s^{N_{0:s-1}}(1)[\eta_s^{N_s} - \Phi_s(\eta_{s-1}^{N_{s-1}})](f)|\mathcal{F}_{s-1}^{N_{0:s-1}}] = 0$$

thus, it follows that

$$\begin{aligned} & \mathbb{E}[[\gamma_{p-2}^{N_{0:p-2}} - \gamma_{p-2}](G_{p-2}(G_{p-1} - 1))[\gamma_{q-2}^{N_{0:q-2}} - \gamma_{q-2}](G_{q-2}(G_{q-1} - 1))] = \\ & \sum_{s=0}^{q-2} \mathbb{E}[\gamma_s^{N_{0:s-1}}(1)^2[\eta_s^{N_s} - \Phi_s(\eta_{s-1}^{N_{s-1}})]^{\otimes 2}(Q_{s,p-2}(\overline{G}_p) \otimes Q_{s,q-2}(\overline{G}_q))]. \end{aligned}$$

Now for any  $n \geq 1$ ,  $f_1, f_2 \in \mathcal{B}_b(E)$ , one can show, using almost the same calculations as above, that the following holds

$$\sum_{s=0}^n \mathbb{E}[\gamma_s^{N_{0:s-1}}(1)^2[\eta_s^{N_s} - \Phi_s(\eta_{s-1}^{N_{s-1}})]^{\otimes 2}(Q_{s,n}(f_1) \otimes Q_{s,n}(f_2))] =$$

$$\mathbb{E}[[\gamma_n^{N_{0:n}} - \gamma_n](f_1)[\gamma_n^{N_{0:n}} - \gamma_n](f_2)].$$

Using this equality with  $n = q - 2$ , and the fact that  $Q_{s,p-2} = Q_{s,q-2}Q_{q-2,p-2}$ , finally

$$\begin{aligned} & \mathbb{E}[[\gamma_{p-2}^{N_{0:p-2}} - \gamma_{p-2}](G_{p-2}(G_{p-1} - 1))[\gamma_{q-2}^{N_{0:q-2}} - \gamma_{q-2}](G_{q-2}(G_{q-1} - 1))] = \\ & \mathbb{E}[[\gamma_{q-2}^{N_{0:n}} - \gamma_{q-2}](Q_{q-2,p-2}(\overline{G}_p))[\gamma_{q-2}^{N_{0:n}} - \gamma_{q-2}](\overline{G}_q)]. \end{aligned}$$

Then, by Lemma 2:

$$\begin{aligned} & \mathbb{E}[[\gamma_{p-2}^{N_{0:p-2}} - \gamma_{p-2}](G_{p-2}(G_{p-1} - 1))[\gamma_{q-2}^{N_{0:q-2}} - \gamma_{q-2}](G_{q-2}(G_{q-1} - 1))] = \\ & \mathbb{E}[(\gamma_{q-2}^{N_{0:q-2}})^{\otimes 2}(Q_{q-2,p-2}(\overline{G}_p) \otimes \overline{G}_q)] - \gamma_{q-2}(1)^2 \eta_{q-2}^{\otimes 2}(Q_{q-2,p-2}(\overline{G}_p) \otimes \overline{G}_q). \end{aligned}$$

Then, one can apply Proposition 1 to obtain that

$$\begin{aligned} & |\mathbb{E}[[\gamma_{p-2}^{N_{0:p-2}} - \gamma_{p-2}](G_{p-2}(G_{p-1} - 1))[\gamma_{q-2}^{N_{0:q-2}} - \gamma_{q-2}](G_{q-2}(G_{q-1} - 1))]| \leq \\ & \frac{C(q-1)\gamma_{q-2}(1)^2}{N_{q-2}} \left\| G_{q-2}(G_{q-1} - 1)Q_{q-2,p-2}(G_{p-2}(G_{p-1} - 1)) \right\|_{\infty}. \end{aligned}$$

□

*Proof of Theorem 2.* Throughout the proof  $C < +\infty$  is a constant whose value may change from line-to-line. It will not depend on the level index. We have

$$\mathbb{E}\left[\left(\frac{\tilde{\gamma}_L^{N_{0:L-2}}(1)}{\gamma_L(1)} - 1\right)^2\right] \leq$$

$$\frac{1}{\gamma_L(1)^2} \mathbb{E}[[\eta_0^{N_0} - \eta_0](G_0)^2] + \frac{1}{\gamma_L(1)^2} \mathbb{E}[(\sum_{p=2}^L [\gamma_{p-2}^{N_{0:p-2}} - \gamma_{p-2}](\overline{G}_p))^2].$$

As  $\gamma_L(1) = Z_L/Z_0 \geq \underline{C}/\overline{C}$  it follows by standard results for i.i.d. random variables that one has

$$\frac{1}{\gamma_L(1)^2} \mathbb{E}[[\eta_0^{N_0} - \eta_0](G_0)^2] \leq \frac{C}{N_0}.$$

Now

$$\begin{aligned} \mathbb{E}[(\sum_{p=2}^L [\gamma_{p-2}^{N_{0:p-2}} - \gamma_{p-2}](\overline{G}_p))^2] &= \sum_{p=2}^L \gamma_{p-2}(1)^2 \mathbb{E}\left[\frac{\gamma_{p-2}^{N_{0:p-2}}(\overline{G}_p)^2}{\gamma_{p-2}(1)^2} - \eta_{p-2}(\overline{G}_p)^2\right] \\ &\quad + 2 \sum_{p=2}^L \sum_{q=2}^{p-1} \mathbb{E}[[\gamma_{p-2}^{N_{0:p-2}} - \gamma_{p-2}](\overline{G}_p)[\gamma_{q-2}^{N_{0:q-2}} - \gamma_{q-2}](\overline{G}_q)]. \end{aligned}$$

Applying Proposition 1 to the terms in the single sum and Lemma 3 to the terms in the double sum, we have that

$$\begin{aligned} \mathbb{E}[(\sum_{p=2}^L [\gamma_{p-2}^{N_{0:p-2}} - \gamma_{p-2}](\overline{G}_p))^2] &\leq C \left( \sum_{p=2}^L \gamma_{p-2}(1)^2 \frac{(p-1)}{N_p} \|\overline{G}_p\|_\infty^2 \right. \\ &\quad \left. + \sum_{p=2}^L \sum_{q=2}^{p-1} \frac{(q-1)\gamma_{q-2}(1)^2}{N_{q-2}} \left\| G_{q-2}(G_{q-1}-1)Q_{q-2,p-2}(G_{p-2}(G_{p-1}-1)) \right\|_\infty \right). \end{aligned}$$

As  $\gamma_{p-2}(1) \leq \overline{C}/\underline{C}$ ,  $\gamma_L(1) \geq \underline{C}/\overline{C}$  one has

$$\frac{1}{\gamma_L(1)^2} \sum_{p=2}^L \gamma_{p-2}(1)^2 \frac{(p-1)}{N_p} \|\overline{G}_p\|_\infty^2 \leq C \sum_{p=2}^L \frac{(p-1)}{N_{p-2}} \|G_{p-1}-1\|_\infty^2.$$

We have

$$\begin{aligned} \frac{1}{\gamma_L(1)^2} \sum_{p=2}^L \sum_{q=2}^{p-1} \frac{(q-1)\gamma_{q-2}(1)^2}{N_{q-2}} \left\| G_{q-2}(G_{q-1}-1)Q_{q-2,p-2}(G_{p-2}(G_{p-1}-1)) \right\|_\infty &= \\ \sum_{p=2}^L \sum_{q=2}^{p-1} \frac{c(q-1)\gamma_{q-2}(1)}{\gamma_L(1)N_{q-2}} \frac{1}{\eta_{q-2}(Q_{q-2,p-2}(1))} \frac{Z_{p-1}}{Z_L} \times & \\ \left\| G_{q-2}(G_{q-1}-1)Q_{q-2,p-2}(G_{p-2}(G_{p-1}-1)) \right\|_\infty. & \end{aligned}$$

Then as  $\gamma_{q-2}(1) \leq \overline{C}/\underline{C}$ ,  $\gamma_L(1) \geq \underline{C}/\overline{C}$ ,  $Z_{p-1} \leq C$ ,  $Z_L \geq C$  and by [7, Lemma 4.1]

$$\frac{Q_{q-2,p-2}(G_{p-2}(G_{p-1}-1))}{\eta_{q-2}(Q_{q-2,p-2}(1))} \leq C \|G_{p-1}-1\|_\infty$$

we have

$$\begin{aligned} \frac{1}{\gamma_L(1)^2} \sum_{p=2}^L \sum_{q=2}^{p-1} \frac{(q-1)\gamma_{q-2}(1)^2}{N_{q-2}} \left\| G_{q-2}(G_{q-1}-1) Q_{q-2,p-2}(G_{p-2}(G_{p-1}-1)) \right\|_{\infty} \leq \\ C \sum_{p=2}^L \sum_{q=2}^{p-1} \frac{(q-1)}{N_{q-2}} \|G_{p-1}-1\|_{\infty} \|G_{q-1}-1\|_{\infty}. \end{aligned}$$

From here one can easily conclude.  $\square$

## References

- [1] ANDRIEU, C., & ROBERTS, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, **37**, 697-725.
- [2] BELLMAN, R. E. (2015) *Adaptive Control Processes: A Guided Tour*. Princeton university press.
- [3] BESKOS, A., CRISAN, D. & JASRA, A. (2014). On the stability of sequential Monte Carlo methods in high dimensions. *Ann. Appl. Probab.*, **24**, 1396–1445.
- [4] BESKOS, A., JASRA, A., LAW, K. J. H, TEMPONE, R. & ZHOU, Y. (2015). Multilevel sequential Monte Carlo samplers. *arXiv preprint arXiv:1503.07259*.
- [5] BICKEL, P., LI, B. & BENGTTSSON, T. (2008). Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the Limits of Contemporary Statistics*, B. Clarke & S. Ghosal, Eds, 318–329, IMS.
- [6] BUNGARTZ, H-J., & GRIEBEL, M. (2004). Sparse grids. *Acta numerica* **13.1**, 147-269.
- [7] CEROU, F., DEL MORAL, P., & GUYADER, A. (2011). A non-asymptotic theorem for unnormalized Feynman-Kac particle models. *Ann. Inst. Henri Poincaré*, **47**, 629–649.
- [8] DEL MORAL, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer: New York.
- [9] DEL MORAL, P., DOUCET, A. & JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, **68**, 411–436.
- [10] GILES, M. B. (2008). Multilevel Monte Carlo path simulation. *Op. Res.* **56**, 607-617.
- [11] GILES, M. B (2015). Multilevel Monte Carlo methods. *Acta Numerica*, **24**, 259-328.

- [12] HEINRICH, S. (2001). Multilevel Monte Carlo methods. *Large-scale scientific computing*. Springer Berlin Heidelberg, 2001. 58-67.
- [13] HOANG, V., SCHWAB, C. & STUART, A. (2013). Complexity analysis of accelerated MCMC methods for Bayesian inversion. *Inverse Prob.*, **29**, 085010.
- [14] HOEL, H., LAW, K. J., & TEMPONE, R. (2015). Multilevel ensemble Kalman filtering. *arXiv preprint arXiv:1502.06069*.
- [15] HOETING, J. A., MADIGAN, D., RAFTERY, A. E., & VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial. *Statist. Sci.*, **14**, 382-401.
- [16] JASRA, A., KAMATANI, K., LAW, K. J., & ZHOU, Y. (2015). Multilevel particle filter. *arXiv preprint arXiv:1510.04977*.
- [17] KETELSEN, C., SCHEICHL, R. & TECKENTRUP, A. L. (2013). A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. *arXiv preprint arXiv:1303.7343*.
- [18] LE MATRE, O. P., & KNIO, O. M. (2010). *Introduction: Uncertainty Quantification and Propagation*. Springer Netherlands.
- [19] ØKSENDAL, B. (2003). *Stochastic Differential Equations*. Springer Berlin Heidelberg.
- [20] RHEE, C. H., & GLYNN, P. W. (2015). Unbiased estimation with square root convergence for SDE models. *Op. Res.*, **63**, 1026–1043.
- [21] SCHWAB, C., & GITTELSO, C. J. (2011). Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. *Acta Numerica*, **20**, 291.
- [22] SCHWEIZER, N. (2012). Non-asymptotic error bounds for sequential MCMC and stability of Feynman-Kac operators. *arXiv preprint arXiv:1204.2382v1*.
- [23] WALSH, J. B. (1986) *An Introduction to Stochastic Partial Differential Equations*. Springer Berlin Heidelberg.
- [24] WALSTROM, J. E., MUELLER, T. D. & MCFARLANE, R. C. (1967). Evaluating uncertainty in engineering calculations. *J. Pet. Tech.* 19.12, 1-595.
- [25] WASSERMAN, L. (2000). Bayesian model selection and model averaging. *J. Math. Psych.*, 44(1), 92-107.
- [26] ZHOU, Y., JOHANSEN, A. M. & ASTON, J. A. D. (2016). Towards automatic model comparison: An adaptive sequential Monte Carlo approach. *J. Comp. Graph. Statist.*, (to appear).